ML Applications:
Text Classification

These slides were assembled by Eric Eaton, with grateful acknowledgement of the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution. Please send comments and corrections to Eric.
Naïve Bayes Review
The Naïve Bayes Graphical Model

- CPTs are estimated via counting
- Laplace smoothing eliminates zero counts:

\[
P(X_j = v \mid Y = y_k) = \frac{1 + c_v}{K + \sum_{v' \in \text{values}(X_j)} c_{v'}}
\]
Example NB Graphical Model

Data:

<table>
<thead>
<tr>
<th>Sky</th>
<th>Temp</th>
<th>Humid</th>
<th>Play?</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>warm</td>
<td>normal</td>
<td>yes</td>
</tr>
<tr>
<td>sunny</td>
<td>warm</td>
<td>high</td>
<td>yes</td>
</tr>
<tr>
<td>rainy</td>
<td>cold</td>
<td>high</td>
<td>no</td>
</tr>
<tr>
<td>sunny</td>
<td>warm</td>
<td>high</td>
<td>yes</td>
</tr>
</tbody>
</table>

| Sky  | Play? | P(Sky | Play) |
|------|-------|--------|
| sunny| yes   | 4/5    |
| rainy| yes   | 1/5    |
| sunny| no    | 1/3    |
| rainy| no    | 2/3    |

| Temp | Play? | P(Temp | Play) |
|------|-------|--------|
| warm | yes   | 4/5    |
| cold | yes   | 1/5    |
| warm | no    | 1/3    |
| cold | no    | 2/3    |

| Humid | Play? | P(Humid | Play) |
|-------|-------|--------|
| high  | yes   | 3/5    |
| norm  | yes   | 2/5    |
| high  | no    | 2/3    |
| norm  | no    | 1/3    |
Example Using NB for Classification

Goal: Predict label for \( x = (\text{rainy}, \text{warm}, \text{normal}) \)

\[
h(x) = \arg\max_{y_k} \log P(Y = y_k) + \sum_{j=1}^{d} \log P(X_j = x_j \mid Y = y_k)
\]
Example Using NB for Classification

Predict label for: \( \mathbf{x} = \text{(rainy, warm, normal)} \)

\[
P(\text{play} | \mathbf{x}) \propto \log P(\text{play}) + \log P(\text{rainy} | \text{play}) + \log P(\text{warm} | \text{play}) + \log P(\text{normal} | \text{play}) \\
\propto \log 3/4 + \log 1/5 + \log 4/5 + \log 2/5 = -1.319 \quad \text{predict PLAY}
\]

\[
P(\neg \text{play} | \mathbf{x}) \propto \log P(\neg \text{play}) + \log P(\text{rainy} | \neg \text{play}) + \log P(\text{warm} | \neg \text{play}) + \log P(\text{normal} | \neg \text{play}) \\
\propto \log 1/4 + \log 2/3 + \log 1/3 + \log 1/3 = -1.732
\]
Document Classification
Document Classification

**PROBLEM SETTING**

Given:
- Representation of a document
- Set of classes $1,...,K$

**Classes:**
- (AI)
  - ML
  - Planning
  - Semantics
  - Garb.Coll.
  - Multimedia
  - GUI

**Training Data:**
- (AI)
  - learning intelligence
  - algorithm
  - reinforcement
  - network...
- (Programming)
  - planning temporal reasoning plan language...
- (HCI)
  - programming semantics language proof...
  - garbage collection memory optimization region...

Based on slide by Chris Manning
Document Classification

**PROBLEM SETTING**

**Given:**
- Representation of a document
- Set of classes $1,...,K$

**Determine:**
- Class to which document $d$ belongs

---

**Classes:**
- (AI)
- (Programming)
- (HCI)
  - ML
  - Planning
  - Semantics
  - Garb.Coll.
  - Multimedia
  - GUI

**Training Data:**
- learning intelligence
- algorithm
- reinforcement network
- planning temporal reasoning
- plan language
- programming semantics
- language proof
- garbage collection
- memory optimization
- region

Based on slide by Chris Manning
Text Classification: Examples

- Add terms to Medline abstracts (e.g. “Conscious Sedation” [E03.250])
- Classify business names by industry
- Classify student essays as A/B/C/D/F
- Classify email as *Spam/Other*
- Classify email to tech staff as *Mac/Windows/…*
- Classify pdf files as *ResearchPaper/Other*
- Determine authorship of documents
- Classify movie reviews as *Favorable/Unfavorable/Neutral*
- Classify technical papers as *Interesting/Uninteresting*
- Classify jokes as *Funny/NotFunny*
- Classify websites of companies by Standard Industrial Classification (SIC) code
Text Classification: Examples

• Best-studied benchmark: *Reuters-21578* newswire stories
  – 9603 train, 3299 test documents, 80-100 words each, 93 classes

ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS
BUENOS AIRES, Feb 26
Argentine grain board figures show crop registrations of grains, oilseeds and their products to February 11, in thousands of tonnes, showing those for future shipments month, 1986/87 total and 1985/86 total to February 12, 1986, in brackets:
• Bread wheat prev 1,655.8, Feb 872.0, March 164.6, total 2,692.4 (4,161.0).
• Maize Mar 48.0, total 48.0 (nil).
• Sorghum nil (nil)
• Oilseed export registrations were:
  • Sunflowerseed total 15.0 (7.9)
  • Soybean May 20.0, total 20.0 (nil)
The board also detailed export registrations for subproducts, as follows....

Categories: grain, wheat (of 93 binary choices)
Spam Filtering

From: "" <takworlld@hotmail.com>
Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY!

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW!

=================================================
Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm
=================================================
Bag of Words Representation

What is the best representation for documents?

Idea: Treat each document as a sequence of words

- Assume that word positions are generated independently

Dictionary: set of all possible words

- Compute over set of documents
- Use Webster’s dictionary, etc.
Bag of Words Representation

Represent document $d$ as a vector of word counts $x$

- $x_j$ represents the count of word $j$ in the document
  - $x$ is sparse (few non-zero entries)
Another View of Naïve Bayes For Document Classification

• Let the model parameters for class $c$ be given by:

$$
\theta_c = \{\theta_{c1}, \theta_{c2}, \ldots, \theta_{c|D|}\}
$$

– $\theta_{cj} = P(\text{word } j \text{ occurs in a document from } c)$
– Also have that $\sum_j \theta_{cj} = 1$

• The likelihood of a document $d$ characterized by $x$ is

$$
P(d \mid \theta_c) = \frac{(\sum_j x_j)!}{\prod_j x_j!} \prod_j (\theta_{cj})^{x_j}
$$

– This is just the multinomial distribution, a generalization of the binomial distribution $\binom{n}{k} p^k (1 - p)^{n-k}$
Another View of Naïve Bayes For Document Classification

• The likelihood of a document $d$ characterized by $x$ is

$$P(d \mid \theta_c) = \frac{(\sum_j x_j)!}{\prod_j x_j!} \prod_j (\theta_{cj})^{x_j}$$

• Use Bayes rule:

$$\log P(\theta_c \mid d) \propto \log \left( P(\theta_c) \prod_{j=1}^{|D|} (\theta_{cj})^{x_j} \right) = \log P(\theta_c) + \sum_{j=1}^{|D|} x_j \log \theta_{cj}$$

Therefore,

$$h(d) = \arg \max_c \left( \log P(\theta_c) + \sum_{j=1}^{|D|} x_j \log \theta_{cj} \right)$$

This is just a linear decision function!
1. Compute dictionary $D$ over training set (if not given)
2. Represent training documents as bags of words over $D$
3. Estimate class priors via counting
4. Estimate conditional probabilities as
   \[
   \hat{\theta}_{cj} = \frac{N_{cj} + 1}{N_c + |D|}
   \]
   - $N_{cj}$ is number of times word $j$ occurs in documents from class $c$
   - $N_c$ is total number of words in all documents from class $c$

- Naïve Bayes model for new documents (represented in $D$) is:
  \[
  h(d) = \arg \max_c \left( \log P(c) + \sum_j x_j \hat{w}_{cj} \right)
  \]
  where $\hat{w}_{cj} = \log \hat{\theta}_{cj}$
What are Some Issues with the Bag of Words Representation?

• Documents have different lengths
• Some words aren’t meaningful to represent the content of a document — e.g., “the”, “a”, etc.
• Rare words may be more meaningful than common words

Need a better representation for the documents...
Eliminate Stop Words

Common, “less-meaningful” words are called stop words

• Delete stop words before doing any document processing

Example stop words:

<table>
<thead>
<tr>
<th>a</th>
<th>because</th>
<th>does</th>
<th>haven't</th>
<th>i</th>
<th>more</th>
<th>our</th>
<th>some</th>
<th>they'll</th>
<th>we'll</th>
<th>why</th>
</tr>
</thead>
<tbody>
<tr>
<td>about</td>
<td>been</td>
<td>doesn't</td>
<td>having</td>
<td>i'd</td>
<td>most</td>
<td>ours</td>
<td>such</td>
<td>they're</td>
<td>we're</td>
<td>why's</td>
</tr>
<tr>
<td>above</td>
<td>before</td>
<td>doing</td>
<td>he</td>
<td>i'll</td>
<td>mustn't</td>
<td>elves</td>
<td>than</td>
<td>that</td>
<td>they've</td>
<td>with</td>
</tr>
<tr>
<td>after</td>
<td>being</td>
<td>don't</td>
<td>he'd</td>
<td>i'm</td>
<td>my</td>
<td>out</td>
<td>that's</td>
<td>this</td>
<td>we've</td>
<td>won't</td>
</tr>
<tr>
<td>again</td>
<td>below</td>
<td>down</td>
<td>he'll</td>
<td>i've</td>
<td>myself</td>
<td>own</td>
<td>those</td>
<td>those</td>
<td>were</td>
<td>would</td>
</tr>
<tr>
<td>against</td>
<td>between</td>
<td>during</td>
<td>he's</td>
<td>if</td>
<td>no</td>
<td>over</td>
<td>than</td>
<td>through</td>
<td>weren't</td>
<td>wouldn't</td>
</tr>
<tr>
<td>all</td>
<td>both</td>
<td>each</td>
<td>her</td>
<td>in</td>
<td>nor</td>
<td>own</td>
<td>to</td>
<td>what</td>
<td>what's</td>
<td>you</td>
</tr>
<tr>
<td>am</td>
<td>but</td>
<td>few</td>
<td>here</td>
<td>into</td>
<td>not</td>
<td>same</td>
<td>too</td>
<td>when</td>
<td>when's</td>
<td>you'd</td>
</tr>
<tr>
<td>an</td>
<td>by</td>
<td>for</td>
<td>here's</td>
<td>is</td>
<td>of</td>
<td>shan't</td>
<td>under</td>
<td>where</td>
<td>where's</td>
<td>you'll</td>
</tr>
<tr>
<td>and</td>
<td>can't</td>
<td>from</td>
<td>hers</td>
<td>isn't</td>
<td>off</td>
<td>she</td>
<td>until</td>
<td>where's</td>
<td>you're</td>
<td>your</td>
</tr>
<tr>
<td>any</td>
<td>cannot</td>
<td>further</td>
<td>herself</td>
<td>it</td>
<td>on</td>
<td>she'd</td>
<td>then</td>
<td>where</td>
<td>you've</td>
<td>yours</td>
</tr>
<tr>
<td>are</td>
<td>could</td>
<td>had</td>
<td>him</td>
<td>it's</td>
<td>once</td>
<td>them</td>
<td>there</td>
<td>which</td>
<td>your</td>
<td>yourselves</td>
</tr>
<tr>
<td>aren't</td>
<td>couldn't</td>
<td>hadn't</td>
<td>himself</td>
<td>its</td>
<td>only</td>
<td>themselves until</td>
<td>there</td>
<td>who</td>
<td>yours</td>
<td></td>
</tr>
<tr>
<td>as</td>
<td>did</td>
<td>has</td>
<td>his</td>
<td>itself</td>
<td>or</td>
<td>these</td>
<td>up</td>
<td>wasn't</td>
<td>who's</td>
<td>yourself</td>
</tr>
<tr>
<td>at</td>
<td>didn't</td>
<td>hasn't</td>
<td>how</td>
<td>let's</td>
<td>other</td>
<td>they</td>
<td>very</td>
<td>was</td>
<td>who</td>
<td>yourselves</td>
</tr>
<tr>
<td>be</td>
<td>do</td>
<td>have</td>
<td>how's</td>
<td>me</td>
<td>ought</td>
<td>they'd</td>
<td>we</td>
<td>we'd</td>
<td>whom</td>
<td></td>
</tr>
</tbody>
</table>


Term Frequency

Term frequency $tf_{t,d}$ is some measure of importance of term $t$ to document $d$

Boolean: $tf_{t,d} = 1$ if $t$ occurs in $d$, 0 otherwise

Raw Counts: $tf_{t,d} = c_{t,d}$
  - $c_{t,d}$ is the number of times $t$ occurs in $d$

Log-Scaled Counts: $tf_{t,d} = \begin{cases} 1 + \log c_{t,d} & \text{if } c_{t,d} > 0 \\ 0 & \text{otherwise} \end{cases}$
  - Reduces relative impact of frequent terms

Normalized Counts: $tf_{t,d} = c_{t,d}/|d|$
  - Normalize raw counts by length of document $|d|$
Inverse Document Frequency

**Idea:** rare terms are more important than common terms

**Example:** if all training documents for a class contain
- the (relatively) common word “water”, and
- the (relatively) rare word “hippopotamus”,
- the term “hippopotamus” is likely more important

**Inverse Document Frequency**

\[ idf_{t,X} = \log \left( \frac{|X|}{|X_t| + 1} \right) \]

- \(X\) is the total set of documents
- \(X_t\) is the subset of documents containing term \(t\)
TF-IDF Transform

• To compensate for issues with raw word counts, use TF-IDF transform on the features with naïve Bayes

\[ tfidf_{t,d,X} = tf_{t,d} \times idf_{t,X} \]

– Represent document as a vector \( x \) of TF-IDF features
– \( x_j \) represents the TF-IDF of word \( j \) in the document

Recommendations:  

(From [Rennie, et al. ICML’03])

– Use raw counts or log-scaled counts for \( tf_{t,d} \)
– Normalize each TF-IDF vector \( x \) to have unit length by \( x = x / \|x\|_2 \) and use these unit vectors in naïve Bayes

You must use the same TF-IDF transform for new documents!

(For more details, see http://people.csail.mit.edu/jrennie/papers/icml03-nb.pdf)
Using SVMs for Document Classification
Words $\rightarrow$ Counts $\rightarrow$ Weight Matrix

<table>
<thead>
<tr>
<th></th>
<th>Antony and Cleopatra</th>
<th>Julius Caesar</th>
<th>The Tempest</th>
<th>Hamlet</th>
<th>Othello</th>
<th>Macbeth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antony</td>
<td>5.25</td>
<td>3.18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.35</td>
</tr>
<tr>
<td>Brutus</td>
<td>1.21</td>
<td>6.1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Caesar</td>
<td>8.59</td>
<td>2.54</td>
<td>0</td>
<td>1.51</td>
<td>0.25</td>
<td>0</td>
</tr>
<tr>
<td>Calpurnia</td>
<td>0</td>
<td>1.54</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cleopatra</td>
<td>2.85</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>mercy</td>
<td>1.51</td>
<td>0</td>
<td>1.9</td>
<td>0.12</td>
<td>5.25</td>
<td>0.88</td>
</tr>
<tr>
<td>worser</td>
<td>1.37</td>
<td>0</td>
<td>0.11</td>
<td>4.15</td>
<td>0.25</td>
<td>1.95</td>
</tr>
</tbody>
</table>

Each document is now represented by a real-valued vector of $|D|$ TF-IDF weights

Based on slide by P. Nayak and P. Raghavan
Documents as Vectors

• So we have a $|D|$-dimensional vector space
  – Terms are axes of the space
  – Documents are points or vectors in this space

• Very high-dimensional:
  – Over 1M words in english
  – More if we allow non-word terms

• Very sparse vectors

• **Idea:** Measure similarity of documents via proximity in the vector space
Why Euclidean Distance is a Bad Idea

- Because Euclidean distance is large for vectors of different lengths

\[ \| q - d_2 \|_2 \] is large, even though the distribution of terms in the query \( q \) and the distribution of terms in the document \( d_2 \) are very similar.
Use Angle Instead of Distance

Thought experiment:
• Take a document \( d \) and append it to itself, creating a new document \( d' \)
• Semantically, \( d \) and \( d' \) have the same content
• But, the Euclidean distance between the two documents can be quite large
• However, note that the angle between the two documents is 0, corresponding to maximal similarity

**Key Idea:** Measure similarity based on angle of vector
From Angles to Cosines

• The following two notions are equivalent:
  – Measure similarity between documents $d_i$ and $d_j$ via decreasing order of the angle between $\mathbf{x}_i$ and $\mathbf{x}_j$
  – Measure similarity in increasing order of cosine($\mathbf{x}_i, \mathbf{x}_j$)

• Cosine is a monotonically decreasing function for the interval $[0^\circ, 180^\circ]$
Length Normalization

• A vector can be (length-) normalized by dividing each of its components by its length (the $L_2$ norm)

\[
x = \frac{x}{||x||_2}
\]

• Dividing a vector by its $L_2$ norm makes it a unit (length) vector (on surface of unit hypersphere)

• Effect on the two documents $d$ and $d'$ ($d$ appended to itself) from earlier slide: they have identical vectors after length-normalization
  
  – Long and short documents now have comparable weights

Based on slide by P. Nayak and P. Raghavan
Cosine Similarity

\( x_i \) and \( x_j \) are TF-IDF weight vectors

\[
\cos(x_i, x_j) = \frac{x_i \cdot x_j}{|x_i||x_j|}
\]

\[
= \frac{x_i}{|x_i|} \cdot \frac{x_j}{|x_j|}
\]

\( \theta \) is the angle between \( x_i \) and \( x_j \)

\[
\cos(x_i, x_j) = \frac{x_i \cdot x_j}{|x_i||x_j|}
\]

\( \cos(x_i, x_j) \) is the cosine similarity of \( x_i \) and \( x_j \)

- Equivalently, the cosine of the angle between \( x_i \) and \( x_j \)
- For unit vectors, cosine similarity is simply the dot product

Based on slide by P. Nayak and P. Raghavan
SVMs for Text Classification

• Use the cosine similarity kernel on TF-IDF features

\[ K(x_i, x_j) = \frac{x_i^\top x_j}{\|x_i\| \cdot \|x_j\|} \]
Advanced Evaluation Metrics
## Confusion Matrix

Given a dataset of $P$ positive instances and $N$ negative instances:

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>TP</td>
<td>FN</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>FP</td>
<td></td>
<td>TN</td>
</tr>
</tbody>
</table>
Accuracy & Error

Given a dataset of $P$ positive instances and $N$ negative instances:

\[
\text{accuracy} = \frac{TP + TN}{P + N}
\]

\[
\text{error} = 1 - \frac{TP + TN}{P + N} = \frac{FP + FN}{P + N}
\]
Why Not Just Use Accuracy?

• How to build a 99.9999% accurate search engine on a low budget....

Search for:  

0 matching results found.

• Users doing information retrieval want to find something and have a certain tolerance for junk
**Precision & Recall**

**Precision**
- the fraction of positive predictions that are correct
- \( P(\text{is pos}|\text{predicted pos}) \)

\[
\text{precision} = \frac{TP}{TP + FP}
\]

**Recall**
- fraction of positive instances that are identified
- \( P(\text{predicted pos}|\text{is pos}) \)

\[
\text{recall} = \frac{TP}{TP + FN}
\]

\[\begin{array}{c|cc}
\text{Actual Class} & \text{Yes} & \text{No} \\
\hline
\text{Yes} & TP & FN \\
\hline
\text{No} & FP & TN \\
\end{array}\]

\[\begin{array}{c|cc}
\text{Actual Class} & \text{Yes} & \text{No} \\
\hline
\text{Yes} & TP & FN \\
\hline
\text{No} & FP & TN \\
\end{array}\]
Receiver Operating Characteristic (ROC)

ROC curves assess predictive behavior

- Originated from signal detection theory
- Common in medical diagnosis, now used for ML

Plots TP rate vs FP Rate

TP rate = TP/P
FP rate = FP/N

Example ROC Plot

True positive rate

False positive rate

Learner L1
Learner L2
Learner L3
Random

Figure and some text provided by Larry Holder
Performance Depends on Threshold

Predict positive if $P(y = 1 \mid x) > \theta$, otherwise negative

- Number of TPs and FPs depend on threshold $\theta$
- As we vary $\theta$, we get different (TPR, FPR) points

Example ROC Plot

Based on slide by Kevin Murphy, Figure by Larry Holder
### ROC Example

<table>
<thead>
<tr>
<th>$i$</th>
<th>$y_i$</th>
<th>$p(y_i = 1 \mid x_i)$</th>
<th>$h(x_i \mid \theta = 0)$</th>
<th>$h(x_i \mid \theta = 0.5)$</th>
<th>$h(x_i \mid \theta = 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.6</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.5</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.4</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$TPR = 5/5 = 1 \quad TPR = 5/5 = 1 \quad TPR = 0/5 = 0$

$FPR = 4/4 = 1 \quad FPR = 0/4 = 0 \quad FPR = 0/4 = 0$
### ROC Example

<table>
<thead>
<tr>
<th>$i$</th>
<th>$y_i$</th>
<th>$p(y_i = 1 \mid x_i)$</th>
<th>$h(x_i \mid \theta = 0)$</th>
<th>$h(x_i \mid \theta = 0.5)$</th>
<th>$h(x_i \mid \theta = 1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.8</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.7</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.6</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td><strong>0.2</strong></td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td><strong>0.6</strong></td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>0.3</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0.2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

$TPR = \frac{5}{5} = 1$  $TPR = \frac{4}{5} = 0.8$  $TPR = 0/5 = 0$

$FPR = \frac{4}{4} = 1$  $FPR = \frac{1}{4} = 0.25$  $FPR = 0/4 = 0$
Receiver Operating Characteristic (ROC)

Example ROC Curve

- True positive rate
- False positive rate

Learner L1
Learner L2
Learner L3
Random

"Ideal" learner

Random model is always diagonal line in ROC space
Receiver Operating Characteristic (ROC)

L1 always dominates L2 and L3
Area Under the ROC Curve

• Can take area under the ROC curve to summarize performance as a single number
  – Be cautious when you see only AUC reported without a ROC curve; AUC can hide performance issues